SLABERT Talk Pretty One Day: Modeling Second Language Acquisition with BERT

Aditya Yadavalli[1] **Alekhya Yadavalli**[2] **Vera Tobin**[2]
[1] Karya Inc.*
[2] Language and Cognition Lab, Case Western Reserve University
aditya@karya.in {alekhya.yadavalli, vera.tobin}@case.edu

Abstract

Second language acquisition (SLA) research has extensively studied cross-linguistic transfer, the influence of linguistic structure of a speaker's native language [L1] on the successful acquisition of a foreign language [L2]. Effects of such transfer can be positive (facilitating acquisition) or negative (impeding acquisition). We find that NLP literature has not given enough attention to the phenomenon of *negative transfer*. To understand patterns of both positive and negative transfer between L1 and L2, we model sequential second language acquisition in LMs. Further, we build a Mutlilingual Age Ordered CHILDES (MAO-CHILDES)—a dataset consisting of 5 typologically diverse languages, i.e., German, French, Polish, Indonesian, and Japanese—to understand the degree to which native Child-Directed Speech (CDS) [L1] can help or conflict with English language acquisition [L2]. To examine the impact of native CDS, we use the TILT-based cross lingual transfer learning approach established by Papadimitriou and Jurafsky (2020) and find that, as in human SLA, language family distance predicts more negative transfer. Additionally, we find that conversational speech data shows greater facilitation for language acquisition than scripted speech data. Our findings call for further research using our novel Transformer-based SLA models and we would like to encourage it by releasing our code, data, and models.

1 Introduction

Cross-linguistic transfer can be described as the influence of native language [L1] properties on a speaker's linguistic performance in a new, foreign language [L2]. The interaction of the linguistic structure of a speaker's L1 with the successful acquisition of L2 results in what are termed as *transfer effects*. Transfer effects appear in various aspects of linguistic performance, including vocabulary, pronunciation, and grammar (Jarvis and Pavlenko, 2007). Cross-linguistic transfer can be positive or negative in nature: positive transfer refers to the facilitating effects of one language in acquiring another (e.g., of Spanish vocabulary in acquiring French) and *negative transfer* between the learner's native [L1] and target [L2] languages, producing errors. The greater the differences between two languages, the greater the negative effects.

While cross-lingual transfer has received considerable attention in NLP research (Wu and Dredze, 2019; Wu et al., 2019; Conneau et al., 2017, 2018; Artetxe et al., 2018; Ruder et al., 2017), most of this research has concentrated on practical implications such as the degree to which the right tokenizer can optimize cross-lingual transfer, and has not looked at the kind of sequential transfer relationships that arise in human second language acquisition. Meanwhile, approaches like the Test for Inductive Bias via Language Model Transfer (TILT) (Papadimitriou and Jurafsky, 2020) focus on positive transfer with divergent pairs of training sets, such as MIDI music and Spanish, to shed light on which kinds of data induce generalizable structural features that linguistic and non-linguistic data share. Patterns of both positive and negative transfer between a given L1 and L2, however, can be a valuable source of information about general processes of second language acquisition and typological relationships between the languages in question (Berzak et al., 2014).

Most cross-lingual models do not mimic how humans acquire language, and modeling the differences between first and second language acquisition is a particularly under-explored area. To engage with questions about second language acquisition using LMs, we model sequential second language acquisition in order to look more closely

* Work done as a visiting researcher at Case Western Reserve University

Code, data and models are here: https://github.com/AdityaYadavalli1/SLABERT

⋆ Authors contributed equally

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 11763–11777

at both positive and negative transfer effects that may occur during the acquisition of L2.

Using Child-Directed Speech (CDS) to create L1 training sets that are naturalistic, ecologically valid, and fine-tuned for language acquisition, we model the kind of cross-linguistic transfer effects that cause linguistic structure of the native L1 to influence L2 language acquisition in our novel Second Language Acquisition BERT (SLABERT) framework. The resulting models, when tested on the BLiMP (Benchmark of Linguistic Minimal Pairs for English) grammar test suite (Warstadt et al., 2020), show that L1 may not only facilitate L2 learning, but can also interfere. To the extent that interference is considered in NLP research, it is often understood simply as a failure of positive transfer in model training. We suggest, instead, that these results should be analyzed in terms of distinctive patterns of both negative and positive transfer, which can reveal not just the existence of generalizable features across datasets, but also finer-grained information about structural features of these languages and their accessibility to second language learners.

2 Related Work

Our work is closely related to and in many ways builds on the work done by Huebner et al. (2021). They proposed that Child-Directed Speech has greater potential than other kinds of linguistic data to provide the structure necessary for language acquisition, and released BabyBERTa, a smaller sized RoBERTa (Liu et al., 2019) model designed to investigate the language acquisition ability of Transformer-based Language Models (TLM) when given the same amount of data as children aged 1-6 get from their surroundings. They also released Zorro, a grammar test suite, that is compatible with the small vocabulary of child-directed input.

Child-directed speech (CDS) refers to the special register adopted by some adults, especially parents, when talking to young children (Saxton, 2009). CDS typically features higher fundamental pitch, exaggerated intonation, slower speech, and longer pauses than Adult-Directed Speech (ADS) (Clark, 2016). Utterances in CDS are usually well-formed grammatically, but are syntactically simpler than ADS, often comprising single word utterances or short declaratives. Adults often repeat words, phrases, and whole utterances in CDS (Küntay and Slobin, 2002; Snow, 1972) and make fewer errors (Broen, 1972) than they do in ADS. CDS also tends to use a smaller and simplified vocabulary, especially with very young children (Hayes and Ahrens, 1988). While the universality and necessity of CDS for language acquisition is a matter of debate (Pinker, 1995; Hornstein et al., 2005; Haggan, 2002), it is likely that the features of CDS are universally beneficial in language acquisition (Saxton, 2009). NLP literature suggests that are certain benefits when models are trained on CDS (Gelderloos et al., 2020). Studies from other fields suggest that the pitch contours, repetitiveness, fluency, and rhythms of CDS make it easier for children to segment speech, acquire constructions, and understand language (Cristia, 2011; Thiessen et al., 2005; Nelson et al., 1986; Ma et al., 2011; Soderstrom et al., 2008; Kirchhoff and Schimmel, 2003). Many of these distinctive qualities of CDS seem tailor-made for human language acquisition, which is why we use CDS data as L1 in our SLABERT models.

Several recent studies confirm that the distinctive distributional features of CDS influence the grammatical and lexical categories that children acquire. For instance, Mintz (2003) found that "frequent frames" in CDS–commonly recurring co-occurance patterns of words in sentences–yield very accurate grammatical category information for both adults and children. Similarly, Veneziano and Parisse (2010) found that patterns of frequent use and, importantly, reinforcement in CDS-specific conversational exchanges were most predictive of the constructions children learn. Together, these findings suggest that both token distribution and the distinctive conversational structure of CDS provide useful reinforcement for acquisition. Therefore, when training our L1 model, we pay attention to qualities of the training input such as the conversational structure.

In second language acquisition (SLA) research, patterns of negative transfer are a topic of much interest and have been considered a source of information both about what happens in second language learning and what it can reveal about the typological relationships between L1 and L2. For instance, Dulay and Burt (1974) show that closely analyzing data from children learning a second language reveals that some errors are due to L1 interference (*negative transfer*), while others arise from developmental cognitive strategies similar to those made during L1 acquisition (*developmental errors*).

Berzak et al. (2014) show a strong correlation between language similarities derived from the structure of English as Second Language (ESL) texts and equivalent similarities obtained directly from the typological features of the native languages. This finding was then leveraged to recover native language typological similarity from ESL texts and perform prediction of typological features in an unsupervised fashion with respect to the target languages, showing that structural transfer in ESL texts can serve as valuable data about typological facts.

The phenomenon of cross-linguistic transfer has received considerable attention in NLP research in the context of multilingual Language Models (Wu and Dredze, 2019; Wu et al., 2019; Conneau et al., 2017, 2018; Artetxe et al., 2018; Ruder et al., 2017). Our investigation is particularly inspired by Papadimitriou and Jurafsky (2020)'s Test for Inductive Bias via Language Model Transfer (TILT). This is a novel transfer mechanism where the model is initially pre-trained on training data [L1]. Next, they freeze a part of the model and fine-tune the model on L2. Finally, they test the resulting model on a test set of L2. We follow a similar approach to our model's second language acquisition.

3 Data

3.1 Why Child-Directed Speech

We wanted L1 training sets that are both realistic and fine-tuned to teach language to developmental (first language) learners. We also wanted to reproduce the findings of Huebner et al. (2021) which suggest that Child-Directed Speech as training data has superior structure-teaching abilities for models compared to scripted adult-directed language.

The BabyBERTa studies (Huebner et al., 2021) found that their LM required less data than RoBERTa to achieve similar (or greater) linguistic/syntactic expertise (as tested by Zorro), and suggested that CDS is better than Wikipedia text for teaching linguistic structure to models. Given these findings and widespread support in cognitive science and linguistics for the facilitative nature of CDS in child language learning, we choose to use CDS data from five different languages as our L1s to examine our hypothesis that preexisting linguistic structure of L1 interacts differentially with the acquisition of L2 (English).

Additionally, building on the Huebner et al. (2021) efforts to find superior training data for LMs in general, we explore the possibility that comparing conversational CDS with scripted ADS is a less fair comparison than comparing the quality of conversational CDS with that of conversational ADS as training input for LMs.

3.1.1 Why CHILDES

Our focus in training the Child-Directed Speech model is on replicating for the LM, as closely as possible, the primary linguistic input of young children. While young children are exposed to passive Adult-Directed Speech, speech that is directed at them and intended to communicate with them plays a more central role in the child's linguistic experience (Soderstrom, 2007). For this reason, we use a language database of naturalistic speech directed at children. The CHILDES (Macwhinney, 2000) database, a component of the larger TalkBank corpus, is a vast repository of transcriptions of spontaneous interactions and conversations between children of varying ages and adults.[1] The database comprises more than 130 corpora from over 40 different languages and includes speech directed at children from ages of 6 months to 7 years. The large selection of languages permits us the necessary flexibility in choosing different languages for our L1 data (see Section 3.1.2 for more on Language Selection). The range of child ages allows us to train our models with increasingly complex linguistic input, emulating the linguistic experience of a growing child.

3.1.2 Language Selection

Our focus is on cross-linguistic transfer of language structure; therefore, we use a simple selection criterion and choose five languages with varying distance from English according to their language family: German, French, Polish, Indonesian, and Japanese. We hypothesize languages that are structurally similar to English should perform better (show more positive transfer and less negative transfer). German, French, and Polish, like English, are all Indo-European languages. However, each of these languages belongs to a unique genus: German and English are Germanic languages, French is a Romance language, and Polish is a Slavic language. While English and French do not share the same genus, there is much overlap between the two languages due to the substantial influence of French on English stretching back to the time of Norman

[1] https://talkbank.org

Language	Vocabulary	Total tokens	Avg. Sentence Length	No. of Children	Utterances
American English	27,723	4,960,141	5.54832	1117	893,989
French	22,809	2,473,989	5.74531	535	487,156
German	59,048	4,795,075	5.65909	134	951,559
Indonesian	21,478	2,122,374	3.97058	9	572,581
Polish	31,462	493,298	5.84276	128	84,578
Japanese	44,789	2,397,386	4.17552	136	588,456
Wikipedia-4	84,231	1,907,706	23.8456	-	80,000
English ADS	55,673	905,378	13.1901	-	74,252

Table 1: MAO-CHILDES corpus statistics: the number of unique tokens, total tokens, the average sentence length, the total number of children, and the mean age of child for each language dataset is presented

Conquest. Japanese belongs to the Japanese language family and Indonesian to the Austronesian language family.

3.1.3 Using the AO-CHILDES corpus

The AO-CHILDES (AO: age-ordered) corpus was created from Huebner and Willits (2021) American English transcripts from the CHILDES database. To curate the American English collection, we followed the same cleaning criteria as Huebner and Willits (2021): only transcripts involving children 0 to 6 years of age were procured, from which child (non-adult) utterances and empty utterances were omitted. The initial CHILDES transcriptions were converted from CHAT transcription format to csv format files using `childes-db` (Sanchez et al., 2019) to conduct the data cleaning processes. The resulting dataset, which now contains 2,000,352 sentences, 27723 unique words, and 4,960,141 total word tokens, forms the American English input. This cleaning process was repeated for the corpora of German, French, Polish, Japanese, and Indonesian to create the dataset for each language (see Table 1 for the language statistics).

3.1.4 MAO-CHILDES

For the sake of simplicity we refer to the corpus resulting from the collective datasets of the six languages as MAO-CHILDES (MAO is short for Multilingual Age-Ordered) to show that the transcripts it contains include a selection of different languages and also are ordered by age of child (see Table 1).

Data in MAO-CHILDES is not uniformly distributed across languages, as seen in Table 1. First, Polish is represented by significantly less data than every other language. Second, Indonesian has a lower number of unique tokens compared to other languages. The Indonesian data is also only collected from conversations with 9 children, a much smaller sample size compared to the other languages, which have sample sizes in the hundreds if not thousands. Third, the average sentence length of the Asian languages—Indonesian and Japanese—is smaller than any of the other languages. The effect of these variations in data, caused by both available resources and natural linguistic characteristics of the languages, on the performance of the cross-lingual model is anticipated.

3.2 Adult-Directed Speech corpus

The Adult-Directed Speech (ADS) corpus comprises conversational speech data and scripted speech data. We build on the BabyBERTa efforts to find superior training data for LMs (in general) by experimenting with conversational ADS and comparing its training utility with that of conversational CDS. This investigation is aimed at narrowing down the true source, child-directed language or conversational language, of the reduced data size requirements of BabyBERTa.

To create our conversational ADS corpus, we use the sample COCA SPOKEN corpus.[2] COCA (Corpus of Contemporary American English) is one of the most widely used corpora of English for its rich representation of texts from a wide range of genres, dialects, and time periods. The SPOKEN genre comprises transcriptions of spontaneous conversations between adults. To clean this sample corpus we followed a three step process:

- All spoken disfluencies such as pauses, laughter, and filler utterances encoded in the spoken transcripts were cleaned.
- All meta tags that mention the names of the speakers were removed.
- Finally, the data was sampled manually to check that the corpus was clean.

[2] https://www.corpusdata.org

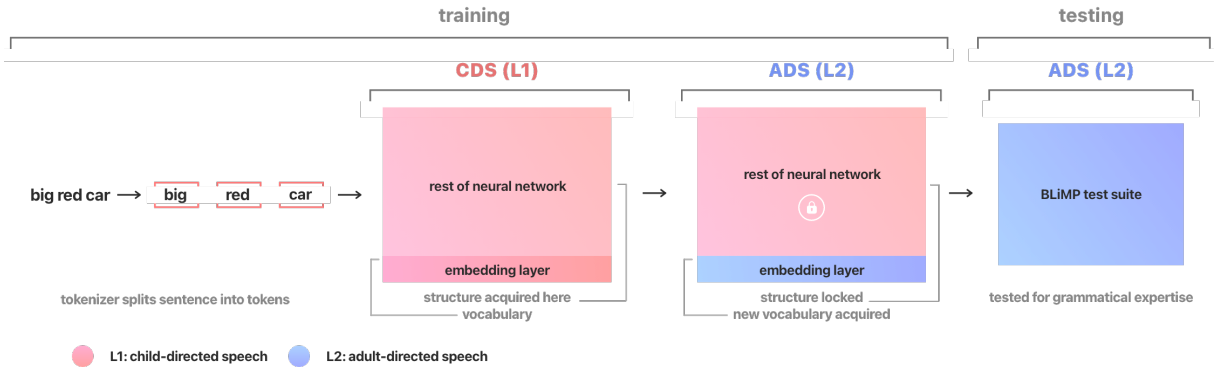

Figure 1: Diagram illustrating our experimental process for each L1, as listed in Table 1. Training occurs in two stages and each model is finally tested on the BLiMP test suite.

After cleaning, we were left with 74,252 utterances. We use this cleaned corpus to train our conversational Adult-Directed Speech (ADS) model.

To replicate the findings of the BabyBERTa study, we also train a model on scripted ADS. To create our scripted ADS corpus, we randomly sample 80,000 sentences from Wikipedia-3 (Huebner et al., 2021), which we term Wikipedia-4, so that the data size of conversational ADS and scripted ADS is approximately equal, to allow fair comparison. All the information about the data we used is in Table 1.

4 Experimental Setup

We use BabyBERTa (Huebner et al., 2021) to run all our experiments. BabyBERTa is a smaller-sized RoBERTa (Liu et al., 2019) tuned to perform well on data of the size of AO-CHILDES. However, we make additional changes to the vocabulary size of the model as we found that to improve the results of the model. The implementation details of the model can be found in Appendix A.1.

We follow the TILT approach introduced by Papadimitriou and Jurafsky (2020) to originally test the LSTM-based (Hochreiter and Schmidhuber, 1997) LM's structure acquisition. Their general approach is followed in the current study with a few notable changes (See Figure 1). Our approach comprises two stages: (1) train the model on L1 (CDS language) (2) freeze all parameters except the word embeddings at the transfer stage of the experiment, and fine-tune the model on L2 (English ADS). Finally, the resulting model is tested on a test set of L2 for which we use the Benchmark of Linguistic Minimal Pairs (BLiMP) (Warstadt et al., 2020), a challenge set for evaluating the linguistic knowledge of the model on major grammatical phenomena in English. Our study deviates from Papadimitriou and Jurafsky (2020) approach in three ways: (1) instead of using LSTM-based LMs we use Transformer-based LMs (Vaswani et al., 2017) (2) they freeze all layers except the word embedding and linear layers between the LSTM layers however, for simplicity we freeze all parameters except the word embeddings (3) while they report their findings based on LM perplexity scores, we use the BLiMP test suite to report how L1 structure (particularly, syntax and semantics) affects L2 acquisition in our Transformer-based LMs.

There are two experiments for which we follow a different procedure than what is explained above:

- In the case of `random-baseline` experiment, we freeze all of the model except the embeddings and let the model train on conversational English ADS. The corresponding tokenizer is also trained on conversational English ADS. This experiment is run in order to have the right benchmark to compare against. This method prevents the model from picking up any grammatical structure from the training data, while allowing it to acquire English vocabulary.

- In the case of the scripted ADS and conversational ADS experiments, we do not employ TILT-based cross lingual transfer. We train the model from scratch on scripted ADS and conversational ADS respectively.

Testing: We use the BLiMP grammar test suite to evaluate the linguistic knowledge of our model. BLiMP consists of 67 paradigms categorized into 12 major grammatical phenomena in English. Each of these 67 datasets comprises 1,000 minimal pairs i.e. pairs of minimally different sentences, one of

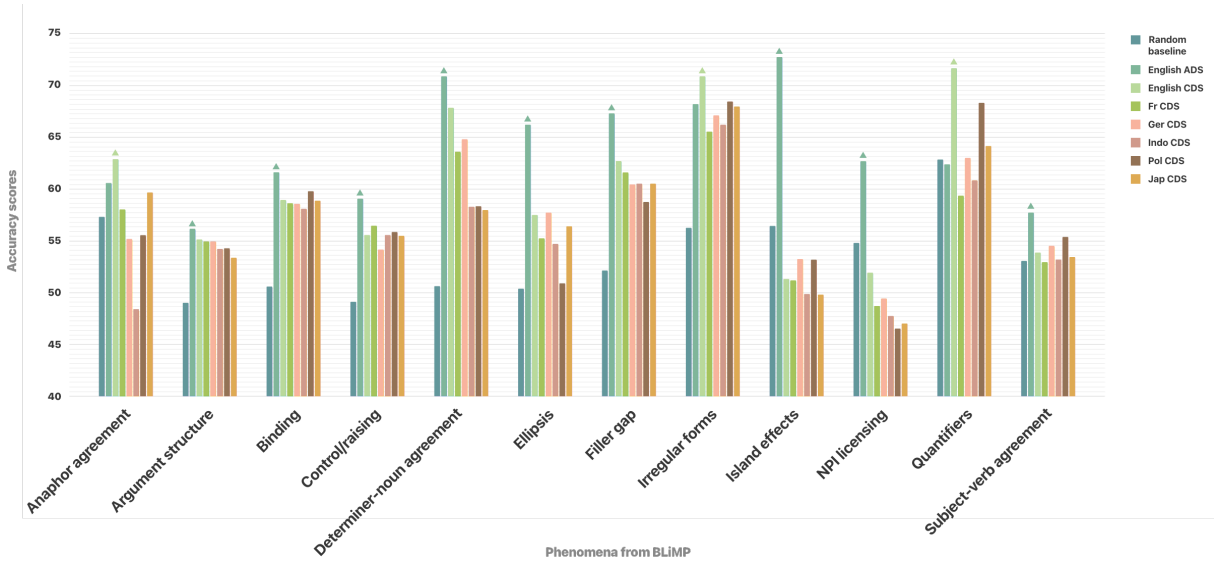

Figure 2: Performance of model on various grammatical phenomena from the BLiMP test suite

which is grammatically acceptable and the other not (refer to Warstadt et al. (2020) for a detailed description of the test suite).

5 Results and Discussion

5.1 Results

The proportion of the BLiMP minimal pairs in which the model assigns a higher probability to the acceptable sentence informs the accuracy of the model. A total of 9 models are compared in their performance using the accuracy scores obtained on 12 different grammatical tests from the BliMP test suite. We report the results for all models in Figure 2 (see Appendix A.2 for detailed results). The model trained on conversational English ADS achieves the highest accuracy and the one trained on Indonesian CDS achieves the lowest. Despite the conversational English ADS corpus size being at least 10x smaller than the CDS corpora sizes, it performs the best in 9 out of 12 grammatical phenomena from the BLiMP test suite. CDS demonstrates higher accuracy only in anaphor agreement, irregular forms, and quantifiers. Overall, English CDS performs 5.13 points behind English ADS. These results show that (conversational) Adult-Directed speech makes for superior training data for models as compared to (conversational) Child-Directed Speech. From Figure 2, we note a few other significant trends:

First, the results indicate that conversational speech data form a superior training data for language models in general as compared to the conventional scripted data. Table 2 compares the performance of models when trained on different types of training inputs of the same language (English): scripted ADS (Wikipedia-4), conversational ADS, and conversational CDS. Among the three, the performance of the model trained on conversational ADS is highest, followed by conversational CDS, and lastly scripted ADS. Important to note here is that, corroborating the findings of the BabyBERTa study, conversational CDS still outperforms scripted ADS (Wikipedia-4) but falls behind compared to conversational ADS. These results suggest that conversational speech data are a more effective training source for models than scripted data (more on this in Section 5.2).

Second, the results show a negative correlation between the distance of the CDS language from English and the performance of the model, i.e., as the typological distance between L1 and L2 increases, the performance of the model decreases. We term this the Language Effect. This finding supports our hypothesis that, given the relation between transfer errors and typological distance between L1 and L2 (Ringbom, 2006), the increasing structural dissimilarities between the L1 (CDS language) and the L2 (always English ADS) should adversely impact the performance of the model (more on this in Section 5.3).

Third, the results show that CDS performs worse than ADS in several grammatical phenomena (9 out of 12). Considering the simplistic and facilitating structure and, more importantly, the ecologically valid nature of CDS, these results engender some

interesting hypotheses which we discuss briefly in Section 5.4.

Fourth, we see several results in which individual models perform poorly on individual tests in ways that are not cleanly predicted by general trends. We believe these results reflect patterns of negative transfer, in which L1-specific structures actively interfere with the acquisition of structures in L2 (more on this in Section 5.5).

5.2 Conversational vs. Scripted Data

The conventional training data for LMs is scripted adult-directed speech, perhaps owing to its easily accessible nature compared to other forms of data, such as conversational ADS or any form of CDS. However, our findings demonstrate that conversational data yields better model performance than scripted data (see Table 2). The best accuracy scores are produced by conversational ADS on 67% of the phenomena, by conversational CDS on 25% of the phenomena, by scripted ADS on 8% of the phenomena. Conversational data may make for a better training input for language acquisition given a higher level of interactive components in its composition which is an essential feature of language acquisition in children. Much of the previous research has looked at what conversational language does for the people who are directly contributing to the conversation in question. For instance, there is a general tendency for speakers to reproduce grammatical (Bock, 1986; Gries, 2005) elements of their interloctor's previous utterances. These behaviors both enhance interactive alignment (Bois, 2014) and ease cognitive load for utterance planning (Bock, 1986; Pickering and Ferreira, 2008). Studies of children's conversational behavior (Veneziano and Parisse, 2010; Köymen and Kyratzis, 2014) show, similarly, that children use their interlocutors' immediately preceding utterances as resources for producing and reinforcing construction types they are in the process of acquiring. Our findings suggest that the resulting distributional patterns of "dialogic syntax" (Bois, 2014) in the conversational record leave a trace that can make conversational data especially informative for model training.

5.3 Language Effect

We selected five languages at varying distances from English according to their language family and examined how structural dissimilarities with increasing distance from English impact the performance of the model. Figure 3 shows the increase in difference between the performance of model trained on English ADS and CDS of the various languages. Our results show negative correlation between the distance of the CDS language from English and the performance of the model, i.e., as the typological distance between L1 and L2 increases, the performance of the model decreases. Based on prior work on transfer errors and typological distance (Ringbom, 2006), this decrease in performance could be the result of negative transfer effects, which tend to increase with increase in typological distance between L1 and L2. Among all CDS languages, English CDS performs closest to English ADS (5.13 points behind ADS), suggesting that even within the same language the linguistic differences between ADS and CDS affect model performance (see Table 2). This is considered as comparisons between other CDS languages and English ADS are made. German shows the next best performance (6.71 points behind English ADS), followed by French (7.27 points behind ADS), Polish (7.57 points behind ADS), Japanese (8.17 points behind ADS), and lastly Indonesian (8.69 points behind ADS). These results confirm our hypothesis that L1s that are structurally closer to L2 (English ADS) perform better, owing to greater degree of positive transfer effects.

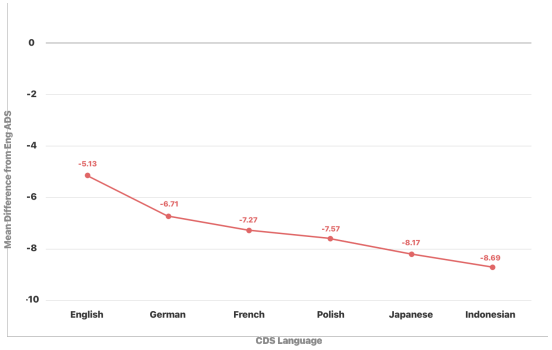

Figure 3: Mean multilingual CDS performance compared to ADS

For human language learners, transfer works both ways: sometimes knowledge of parallel structures in the native language facilitate performance in the new language. Other times, there is interference from the native language, resulting in errors. The SLABERT models, similarly, show evidence of both positive and negative transfer. As with human second-language learners, some of the errors we see in SLABERT performance suggest the effect of negative transfer from native [L1] language,

Phenomenon	Wikipedia-4	Conversational ADS	Conversational CDS
Ananaphor Agreement	51.4	60.6	62.9
Argument Structure	54.5	56.1	55.1
Binding	60.7	61.6	58.9
Control/Raising	48.8	59.1	55.6
Determiner Noun Agreement	65.2	70.9	67.8
Ellipses	68.6	66.2	57.5
Filler Gap	62.4	67.3	62.6
Irregular Forms	61.8	68.2	70.9
Island Effects	51.8	72.7	51.3
NPI Licensing	53.7	62.6	51.9
Quantifiers	58.5	62.4	71.7
Subject Verb Agreement	54.9	57.7	53.8

Table 2: Performance of model on BLiMP test suite when trained on different types of input data.

while others can be characterized as developmental, in that they are similar to the kinds of errors that even native human speakers will make on their way to learning the target constructions.

5.4 CDS & Sources of Errors in Language Learning

Our results show that CDS performs worse than ADS in a majority (9 out of 12) of the grammatical phenomena from the BLiMP test suite (see Figure 2). We discuss some theoretical explanations for these results.

Negation and NPIs: Child language acquisition research strongly suggests that mastering the full range of negative licensing and anti-licensing contexts takes a long time. Across languages, detailed acquisition studies find that children do use NPIs with licensing expressions consistently by age 3 or 4 (Tieu, 2013; Lin et al., 2015) but only with a limited range of negative licensers. Moreover, Schwab et al. (2021) showed that, even 11 and 12-year-olds, whose language input by that age is entirely ADS, are still in the process of learning some polarity-sensitive expressions. Thus, CDS input alone may not be sufficient for learning the licensing conditions for NPIs. Previous NLP literature also suggests that negation is particularly challenging for language models to learn (Kassner and Schütze, 2019; Ettinger, 2019). Given this, and acquisition studies that have shown that learning licensing conditions for NPIs goes hand-in-hand with learning negation (van der Wal, 1996), we expected our model trained on CDS to make *developmental errors* on tests related to NPIs. As discussed in Section 5.5, as a Slavic language, Polish also has distinctive constraints on the appearance of NPIs that are the result of competition with grammatical constraints not present in English. In this case, NPI performance is likely subject to both *developmental* errors and *negative transfer* .

Longer Distance Dependencies: Short and simple sentences are characteristic of CDS. However, it is likely that such utterances do not make ideal training input for LMs to learn long-distance dependencies (LDDs). Consequently, we expect all models trained on CDS data to be negatively impacted on tests that demand long-distance dependency understanding. Island effects, the phenomenon that showed the widest difference in performance compared to ADS-trained (-21.3 points), is one such phenomenon in the BLiMP test suite, requiring long-distance dependency understanding to perform well (Sprouse and Hornstein, 2013). Ellipsis and filler-gap structures also depend on LDDs and also suffer from significant decreases in scores compared to ADS (-10.8 and -6.5 points, respectively). This also applies to binding and control/raising phenomena (-2.8 and -3.6 respectively); however, island effects, ellipsis, and filler-gap tests are particularly affected by the model's lack of LDD understanding.

Phenomena That Confuse Humans: Warstadt et al. (2020) report human performance scores which we use to gain an understanding of how our model performs on tests compared to humans. From the reported human performance scores, we observe that not all of the grammatical phenomena in the BLiMP test suite are equally transparent to humans. Human performance on 8 out of 12 phenomena is below 90 points and 3 of those are below 85 points. The lowest is a mean score of 81 for tests on argument structure, where the CDS-trained and ADS-trained models are also seen struggling (rather more seriously) with a mean score of 55.1 and 56.1, respectively. For control/raising, similarly, human performance has a mean score of 84

points while CDS-trained and ADS-trained models have mean scores of 55.6 and 59.1 respectively. We expect CDS to perform poorly on these tests, which are challenging even for people.

5.5 Negative Transfer

There are tests where performance of CDS-trained models would be expected to be better given the nature of the phenomena and the characteristics of CDS utterances. However, CDS underperforms compared to ADS even on tests we might expect to be in its wheelhouse. In particular, determiner-noun agreement and subject-verb agreement are the kinds of phenomena that should be easy for the model to learn even from shorter utterances and with relatively small vocabulary size, since they are matters of simple, regular morphology. The results, therefore, are interesting. We hypothesize one reason we do not see good transfer boosts from other-language CDS on these is that patterns of morphology are very language specific.

Looking broadly at the performance of non-English CDS models, we suggest that these results reflect negative cross-linguistic transfer. For example, the distribution of negative polarity items in Polish and many other Slavic languages displays what has been termed the "Bagel problem" (Pereltsvaig, 2006): because of conflicts with the demands of strict negative concord (in which negation requires multiple elements of an expression must all appear in their negative forms), in Slavic languages, there are NPIs that never appear in what would otherwise be the canonical context of negative polarity licensing, i.e. direct negation (Hoeksema, 2012). In this way, language-specific paradigmatic patterns supersede the general correlational relationship between NPIs and their licensing contexts, producing an opportunity for *negative transfer* and L1 interference effects.

6 Conclusion

In this paper, we explore how second language acquisition research and models of second language acquisition can contribute to questions in NLP about the learnability of grammar. Drawing from the previous research on the unique role of child-directed speech (CDS) in language acquisition, we investigate the potential of spontaneously generated CDS to form a special source from which LMs can acquire the structure necessary for first language acquisition. To test sequential second language acquisition in LMs, we introduce SLABERT. The results from our experiments suggest that while positive transfer is a lot more common than negative transfer, negative transfer occurs in LMs just like it occurs in English Second Language (ESL) learners. We believe these novel findings call for further research on this front, and suggest that models like SLABERT can provide useful data for testing questions about both language acquisition and typological relationships through patterns of cross-linguistic transfer. To support this, we release our code, novel MAO-CHILDES corpus, and models.

7 Limitations

Given that many special properties of Child-Directed Speech are not present in text, we would have liked to work on a multimodal dataset, where both visual and speech information would be present. More specifically, we would have liked to test the effect of the following:

- Grounding the language models in vision to test the effect of joint attention (Rowe, 2012; Akhtar and Gernsbacher, 2007). Joint attention refers to the phenomena where the caregiver's and the child's coordinated attention to each other to a third object or an event.

- Child-Directed Speech is known to have special prosodic properties such as higher variability in pitch (Fernald et al., 1989; McRoberts and Best, 1997; Papousek et al., 1991), lengthening of vowels and pauses (Albin and Echols, 1996; Ratner, 1986; Fernald et al., 1989), context-specific intonational contours (Katz et al., 1996; Papousek et al., 1991; Stern et al., 1982). These properties have been suggested by many researchers to serve as a mechanism for getting the infants attention (Cruttenden, 1994; Ferguson, 1977; Fernald, 1989). This attentive role may be considered to be beneficial for language development in children (Garnica, 1977). As our models only take text as the input, we were unable to test the relationship the between these properties and language acquisition in neural network based models have.

- Caregivers give a lot of feedback when young children are first producing and acquiring language (Soderstrom, 2007). Our current mainstream language models are not interactive.

Therefore, it is difficult to incorporate the feedback loop and the test the effect of the same in models' language acquisition.

As it is, our findings suggest that many of the most important facilitative features of Child-Directed Speech are relevant to precisely those formal and conceptual aspects of language acquisition that are not captured by text-based language models.

In this paper, we have tested the effect of native CDS in L2 acquisition with 5 typologically diverse languages. However, there is enormous scope to test the effect of the same with many more different languages, which may lead to more pointed implications and conclusions than the findings offered here.

8 Ethics Statement

We use publicly available CHILDES data to build our corpora (MAO-CHILDES). Please read more about their terms before using the data.[3] We use the dataset extracted from the CHILDES database only for research purposes and not for commercial reasons. We will release the dataset upon publication under the same license as CHILDES and this is compatible with the license of CHILDES database (Macwhinney, 2000). The results of this study are reported on a single run as part of measures taken to avoid computation wastage. We do not foresee any harmful uses of this work.

Acknowledgements

We would like to acknowledge Philip Huebner for clearing our queries regarding the BabyBERTa code-base. We would also like to thank Saujas Vaduguru for helping us improve our initial drafts. We also thank the anonymous reviewers for their feedback on our work. This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University.

References

Nameera Akhtar and Morton Ann Gernsbacher. 2007. Joint attention and vocabulary development: A critical look. *Language and Linguistics Compass*, 1 3:195–207.

Drema Dial Albin and Catharine H. Echols. 1996. Stressed and word-final syllables in infant-directed speech. *Infant Behavior & Development*, 19:401–418.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2014. Reconstructing native language typology from foreign language usage. In *CoNLL*.

J. Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18:355–387.

John W. Du Bois. 2014. Towards a dialogic syntax. *Cognitive Linguistics*, 25:359–410.

Patricia Broen. 1972. The verbal environment of the language-learning child. *Monographs of the American Speech and Hearing Association*, 17.

Eve V. Clark. 2016. *First Language Acquisition*, 3 edition. Cambridge University Press.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv'e J'egou. 2017. Word translation without parallel data. *ArXiv*, abs/1710.04087.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alejandrina Cristia. 2011. Fine-grained variation in caregivers' /s/ predicts their infants' /s/ category. *The Journal of the Acoustical Society of America*, 129 5:3271–80.

Alan Cruttenden. 1994. *Phonetic and prosodic aspects of Baby Talk*, page 135–152. Cambridge University Press.

Heidi C. Dulay and Marina K. Burt. 1974. Errors and strategies in child second language acquisition. *TESOL Quarterly*, 8:129.

Allyson Ettinger. 2019. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Ch A. Ferguson. 1977. Baby talk as a simplified register snow. In Catherine E. Snow and Charles A. Ferguson, editors, *Talking to Children*, pages 209–235. Cambridge University Press.

[3] https://talkbank.org

Anne Fernald. 1989. Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development*, 60(6):1497–1510.

Anne Fernald, Traute Taeschner, Judy Dunn, Mechthild Papousek, Bénédicte de Boysson-Bardies, and I Fukui. 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, 16:477 – 501.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Olga K. Garnica. 1977. Some prosodic and paralinguistic features of speech to young children. In Catherine E. Snow and Charles A. Ferguson, editors, *Talking to Children*, pages 63–88. Cambridge University Press.

Lieke Gelderloos, Grzegorz Chrupała, and A. Alishahi. 2020. Learning to understand child-directed and adult-directed speech. In *Annual Meeting of the Association for Computational Linguistics*.

Stefan Th. Gries. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, 34:365–399.

Madeline Haggan. 2002. Self-reports and self-delusion regarding the use of motherese: implications from kuwaiti adults. *Language Sciences*, 24(1):17–28.

Donald P. Hayes and Margaret G. Ahrens. 1988. Vocabulary simplification for children: a special case of 'motherese'? *Journal of Child Language*, 15(2):395–410.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Jack Hoeksema. 2012. On the natural history of negative polarity items. *Linguistic Analysis*, 44:3–33.

Norbert Hornstein, Jairo Nunes, and Kleanthes K. Grohmann. 2005. *Understanding Minimalism*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Philip A. Huebner and Jon A. Willits. 2021. Chapter eight - using lexical context to discover the noun category: Younger children have it easier. In Kara D. Federmeier and Lili Sahakyan, editors, *The Context of Cognition: Emerging Perspectives*, volume 75 of *Psychology of Learning and Motivation*, pages 279–331. Academic Press.

Scott Jarvis and Aneta Pavlenko. 2007. Crosslinguistic influence in language and cognition.

Nora Kassner and Hinrich Schütze. 2019. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Annual Meeting of the Association for Computational Linguistics*.

Gary S. Katz, Jeffrey F. Cohn, and Christopher A. Moore. 1996. A combination of vocal fo dynamic and summary features discriminates between three pragmatic categories of infant-directed speech. *Child development*, 67 1:205–17.

Katrin Kirchhoff and Steven M. Schimmel. 2003. Statistical properties of infant-directed versus adult-directed speech: insights from speech recognition. *The Journal of the Acoustical Society of America*, 117 4 Pt 1:2238–46.

Bahar Köymen and Amy Kyratzis. 2014. Dialogic syntax and complement constructions in toddlers' peer interactions. *Cognitive Linguistics*, 25:497 – 521.

Aylin Küntay and Dan Slobin. 2002. Putting interaction back into child language: Examples from turkish. *Psychology of Language and Communication, v.6 (2002)*, 6.

Jing Lin, F. P. Weerman, and Hedde Zeijlstra. 2015. Emerging npis: The acquisition of dutch hoeven 'need'. *The Linguistic Review*, 32:333 – 374.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Weiyi Ma, Roberta Michnick Golinkoff, Derek M. Houston, and Kathy Hirsh-Pasek. 2011. Word learning in infant- and adult-directed speech. *Language Learning and Development*, 7:185 – 201.

Brian Macwhinney. 2000. The childes project: Tools for analyzing talk (third edition): Volume i: Transcription format and programs, volume ii: The database. *Computational Linguistics - COLI*, 26:657–657.

Gerald McRoberts and Catherine T. Best. 1997. Accommodation in mean f0 during mother–infant and father–infant vocal interactions: a longitudinal case study. *Journal of Child Language*, 24:719 – 736.

Toben H. Mintz. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90:91–117.

Deborah G Kemler Nelson, Kathy Hirsh-Pasek, Peter W. Jusczyk, and Kimberly Wright Cassidy. 1986. How the prosodic cues in motherese might assist language learning. *Journal of Child Language*, 16:55 – 68.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, Online. Association for Computational Linguistics.

Mechthild Papousek, Hanŭs Papousek, and David T Symmes. 1991. The meanings of melodies in motherese in tone and stress languages. *Infant Behavior & Development*, 14:415–440.

Asya Pereltsvaig. 2006. Small nominals. *Natural Language and Linguistic Theory*, 24:433–500.

Martin John Pickering and Victor S. Ferreira. 2008. Structural priming: a critical review. *Psychological bulletin*, 134 3:427–59.

Steven Pinker. 1995. *The Language Instinct*. PENGUIN.

Nan Bernstein Ratner. 1986. Durational cues which mark clause boundaries in mother–child speech. *Journal of Phonetics*, 14:303–309.

Håkan Ringbom. 2006. *Cross-linguistic Similarity in Foreign Language Learning*. Multilingual Matters, Bristol, Blue Ridge Summit.

Meredith L. Rowe. 2012. A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child development*, 83 5:1762–74.

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2017. A survey of cross-lingual word embedding models. *J. Artif. Intell. Res.*, 65:569–631.

Alessandro Sanchez, Stephan Meylan, Mika Braginsky, Kyle Macdonald, Daniel Yurovsky, and Michael Frank. 2019. childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, 51.

Matthew L. Saxton. 2009. The inevitability of child directed speech.

Juliane Schwab, Mingya Liu, and Jutta L. Mueller. 2021. On the acquisition of polarity items: 11- to 12-year-olds' comprehension of german npis and ppis. *Journal of Psycholinguistic Research*, 50:1487 – 1509.

Catherine E. Snow. 1972. Mothers' speech to children learning language. *Child Development*, 43(2):549–565.

Melanie Soderstrom. 2007. Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27:501–532.

Melanie Soderstrom, Megan Stratton Blossom, Rina Foygel, and James L. Morgan. 2008. Acoustical cues and grammatical units in speech to two preverbal infants*. *Journal of Child Language*, 35:869 – 902.

Jon Sprouse and Norbert Hornstein. 2013. *Experimental syntax and island effects: Toward a comprehensive theory of islands*, page 1–18. Cambridge University Press.

Daniel N. Stern, Susan J. Spieker, and K. Mackain. 1982. Intonation contours as signals in maternal speech to prelinguistic infants. *Developmental Psychology*, 18:727–735.

Erik D. Thiessen, Emily A. Hill, and Jenny R. Saffran. 2005. Infant-directed speech facilitates word segmentation. *Infancy*, 7(1):53–71.

Lyn Tieu. 2013. Logic and grammar in child language: How children acquire the semantics of polarity sensitivity.

S. van der Wal. 1996. Negative polarity items and negation: Tandem acquisition.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Edy Veneziano and Christophe Parisse. 2010. The acquisition of early verbs in french: Assessing the role of conversation and of child-directed speech. *International Conference on Infant Studies 2010*, 30.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *ArXiv*, abs/1911.01464.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

A Appendix

A.1 Implementation Details

We conduct our experiments using BabyBERTa (Huebner et al., 2021), a RoBERTa-based model (Liu et al., 2019), with 8 hidden layers, 8 attention heads, and dimensionality of the encoder layer being 256, dimensionality of the intermediate or the feed-forward layer in the transfer based encoder being 1024. We train this model with a learning rate

Phenomenon	Paradigm	Baseline	Eng ADS	Wikipedia	Eng CDS	Ger CDS	Fr CDS	Pol CDS	Jap CDS	Indo CDS	Acceptable Example	Unacceptable Example
ANAPHOR AGREEMENT	anaphor_gender_agreement	59.1	55.6	46.6	60.4	47.2	51.6	53.6	55.2	40.0	Karla could listen to **herself**.	Karla could listen to **himself**.
	anaphor_number_agreement	53.9	65.5	56.1	65.3	63.1	64.4	57.3	64.1	56.6	Susan revealed **herself**.	Susan revealed **themselves**.
ARGUMENT STRUCTURE	animate_subject_passive	52.2	65.1	57.5	65.1	64.2	59.7	67.3	60.1	57.3	Lisa was kissed by the **boys**.	Lisa was kissed by the **blouses**.
	animate_subject_trans	37.8	56.2	60.8	57.2	59.6	62.1	60.7	58.1	60.3	**Most banks** have praised Raymond.	**The jackets** have praised Raymond.
	causative	51.0	54.7	58.5	54.1	58.0	58.6	56.0	54.9	55.9	April had dropped the truck.	April had existed the truck.
	drop_argument	43.5	43.5	42.9	43.3	45.5	42.3	42.6	39.4	44.4	Noah approached.	Noah works with.
	inchoative	42.4	52.9	46.5	51.7	49.1	48.4	43.9	47.1	47.5	A lot of closets could fling open.	A lot of closets could buy.
	intransitive	37.0	41.0	39.9	39.8	42.1	38.3	39.1	37.0	39.5	Some guests hadn't left.	Some guests hadn't boasted about.
	passive_1	56.5	65.0	59.3	59.7	58.7	62.8	61.7	61.6	61.0	Tracy isn't fired by Jodi's daughter.	Tracy isn't muttered by Jodi's daughter.
	passive_2	55.7	63.9	61.4	62.3	58.6	61.5	59.8	62.1	59.2	The Clintons were attacked.	The Clintons were replied.
	transitive	58.6	62.8	63.8	62.5	58.6	60.7	56.6	59.7	62.2	Some turtles **alarm** Kimberley.	Some turtles **come here** Kimberley.
BINDING	principle_A_c_command	45.0	51.2	56.1	48.0	54.9	46.9	53.1	54.7	53.9	A guy that has seen the wheelbarrow notices **himself**.	A guy that has seen the wheelbarrow notices **itself**.
	principle_A_case_1	75.0	98.4	94.6	97.5	97.3	99.0	97.4	97.3	98.6	Angela wasn't thinking that **she** likes Susan.	Angela wasn't thinking that **herself** likes Susan.
	principle_A_case_2	49.8	64.1	71.4	54.6	54.6	56.4	54.6	53.1	46.3	Eric imagines himself **taking** every rug.	Eric imagines himself **took** every rug.
	principle_A_domain_1	65.1	77.8	80.5	77.7	75.0	76.5	76.4	74.1	74.1	Carl imagines that Maria does leave **him**.	Carl imagines that Maria does leave **himself**.
	principle_A_domain_2	49.7	48.9	58.0	50.1	53.1	53.6	51.4	48.9	54.7	Gary imagined most reports upset **themselves**.	Gary imagined most reports upset **himself**.
	principle_A_domain_3	49.5	51.5	48.2	50.2	49.7	50.0	51.2	49.8	50.7	**Gina explains Alan** fires himself.	**Alan explains Gina** fires himself
	principle_A_reconstruction	46.7	39.4	16.2	34.1	25.2	27.5	34.4	34.0	27.9	It's herself that Andrea attacked.	It's herself that attacked Andrea.
CONTROL/ RAISING	existential_there_object_raising	62.7	61.6	53.2	60.1	61.5	62.9	63.6	66.4	64.7	**Mary can declare** there to be some ladders falling.	**Mary can entreat** there to be some ladders falling.
	existential_there_subject_raising	55.9	57.9	39.7	50.9	46.8	56.2	50.7	45.0	50.0	There is soon to be a cat existing.	There is willing to be a cat existing.
	expletive_it_object_raising	63.0	57.5	52.8	63.9	63.3	65.1	63.3	64.8	64.4	**Nina anticipated it to be** nice that Jacqueline exists.	**Nina obligated it to be** nice that Jacqueline exists.
	tough_vs_raising_1	53.5	55.6	60.4	56.0	54.7	50.4	58.6	56.1	52.1	James is pleasant to flee from.	James is apt to flee from.
	tough_vs_raising_2	50.1	62.8	38.0	46.9	44.2	47.6	42.8	44.8	46.3	Thomas isn't sure to hug Margaret.	Thomas isn't annoying to hug Margaret.
DETERMINER-NOUN AGR.	determiner_noun_agreement_1	53.1	74.8	67.8	74.9	72.1	68.1	61.3	62.3	65.3	Raymond is selling this **sketch**.	Raymond is selling this **sketches**.
	determiner_noun_agreement_2	52.8	77.1	73.4	70.3	73.5	70.6	62.2	63.1	68.0	Craig had cared for that **dancer**.	Craig had cared for those **dancer**.
	determiner_noun_agreement_irregular_1	49.1	66.1	60.6	63.8	65.2	60.2	56.6	56.6	58.4	Laurie hasn't lifted those **cacti**.	Laurie hasn't lifted those **cactus**.
	determiner_noun_agreement_irregular_2	52.0	75.6	73.4	75.8	73.8	70.2	65.1	68.1	69.8	All boys boast about that **child**.	All boys boast about that **child**.
	determiner_noun_agreement_with_adj_1	52.4	70.4	64.7	63.8	60.5	59.9	53.5	53.3	51.8	Sara hasn't bored this displeased **senator**.	Sara hasn't bored this displeased **senators**.
	determiner_noun_agreement_with_adj_2	49.7	67.6	58.7	67.2	55.4	59.7	56.1	53.9	49.2	Cynthia scans these hard **books**.	Cynthia scans this hard **books**.
	determiner_noun_agreement_with_adj_irregular_1	52.1	66.5	64.5	62.7	60.7	61.5	56.1	51.8	52.6	Heidi returns to that big **woman**.	Heidi returns to that big **women**.
	determiner_noun_agreement_with_adj_irregular_2	50.2	68.9	58.8	63.9	56.8	58.2	55.6	54.3	50.8	Julia questioned those small **children**.	Julia questioned those small **children**.
ELLIPSIS	ellipsis_n_bar_1	49.4	71.6	65.5	50.3	45.5	43.9	41.1	47.4	38.4	That book bored many troubled sons and Theresa bored few.	That book bored many sons and Theresa bored few troubled.
	ellipsis_n_bar_2	45.2	60.8	71.6	64.7	69.9	66.5	60.6	65.3	70.9	Renee helps one girlfriend and Roger helps a few ill girlfriend.	Renee helps one unlucky girlfriend and Roger helps a few ill.
FILLER GAP	wh_questions_object_gap	65.5	62.8	60.1	65.6	60.1	58.3	54.7	58.5	58.3	Teresa knew that man that April remembered.	Teresa knew who April remembered that man.
	wh_questions_subject_gap	68.9	72.7	87.3	78.6	83.4	82.8	75.4	82.9	82.1	Tammy notices a coat that hurt Tiffany.	Tammy notices who a coat hurt Tiffany.
	wh_questions_subject_gap_long_distance	64.4	85.6	91.9	87.9	88.2	86.9	82.0	84.1	85.1	Regina sees that candle that Steve lifts that might impress every doctor.	Regina sees who that candle that Steve lifts might impress every doctor.
	wh_vs_that_no_gap	67.4	83.3	95.5	83.6	81.2	82.3	80.6	83.2	86.9	Sandra was figuring out that these guys saw Douglas.	Sandra was figuring out that these guys saw Douglas.
	wh_vs_that_no_gap_long_distance	66.2	87.6	93.6	88.2	81.3	84.8	77.3	85.0	85.1	Ann does remember that those skirts that annoyed Kayla stun Dan.	Ann does remember who those skirts that annoyed Kayla stun Dan.
	wh_vs_that_with_gap	28.5	62.8	3.5	20.9	14.5	17.9	19.5	13.3	13.9	Nina has learned who most men sound like.	Nina has learned that most men sound like.
	wh_vs_that_with_gap_long_distance	30.9	16.1	5.2	13.7	14.3	17.9	21.7	16.5	12.1	A lady has remembered who the actors conceal.	A lady has remembered that the actors conceal.
IRREGULAR FORMS	irregular_past_participle_adjectives	35.9	58.4	73.1	68.8	57.6	65.1	66.2	62.8	62.9	The **hidden** bicycles weren't exposed.	The **hid** bicycles weren't exposed.
	irregular_past_participle_verbs	61.4	77.9	50.5	72.9	76.5	66.0	70.7	73.1	69.4	The mushroom **went** bad.	The mushroom **gone** bad.
ISLAND EFFECTS	adjunct_island	52.4	77.8	53.3	53.3	50.5	50.8	45.5	52.4	54.1	Who should Derek hug after shocking Richard?	Who should Derek hug Richard after shocking?
	complex_NP_island	52.0	71.6	51.7	50.2	47.0	46.9	52.7	46.8	46.5	What can't a guest who would like some actor argue about?	What can't some actor argue about a guest who would like?
	coordinate_structure_constraint_complex_left_branch	42.6	66.2	25.9	29.9	36.3	32.8	33.7	25.7	30.3	Which teenagers had Tamara hired and Grace fired?	Which had Tamara hired teenagers and Grace fired?
	coordinate_structure_constraint_object_extraction	48.7	74.8	66.9	46.3	45.8	41.8	46.1	48.9	41.4	Who were all men and Eric leaving?	Who were all men leaving and Eric?
	left_branch_island_echo_question	70.8	63.3	82.9	61.8	78.7	74.0	75.9	69.9	78.0	**Irene had messed up whose** rug?	**Whose had Irene messed up** rug?
	left_branch_island_simple_question	39.4	87.6	42.4	52.5	53.6	49.6	53.1	40.9	44.6	Whose museums had Dana alarmed?	Whose had Dana alarmed museums?
	sentential_subject_island	49.7	54.9	47.2	54.1	50.7	50.5	50.7	50.8	43.4	Who had the patients' cleaning those banks upset.	Who had the patients' cleaning upset those banks.
	wh_island	69.4	85.6	44.1	62.0	63.1	63.0	67.4	62.7	60.3	Who have those men revealed **they** helped?	Who have those men revealed **who** helped?
NPI LICENSING	matrix_question_npi_licensor_present	27.7	26.7	13.3	24.1	17.2	23.8	19.9	23.5	21.7	**Had Bruce** ever played?	**Bruce had** ever played.
	npi_present_1	36.3	74.8	56.0	35.4	41.7	48.3	42.3	37.9	36.1	Even Suzanne has **really** joked around.	Even Suzanne has **ever** joked around.
	npi_present_2	34.1	48.2	61.5	30.2	43.4	46.7	40.1	39.8	36.0	Tamara **really** exited those mountains.	Tamara **ever** exited those mountains.
	only_npi_licensor_present	78.2	83.3	87.6	58.3	51.3	41.5	36.6	49.0	44.9	**Only Bill would** ever complain.	**Even Bill would** ever complain.
	only_npi_scope	59.6	57.4	34.4	72.3	50.6	56.3	56.3	48.9	61.3	**Only a popsicle that Danielle admires** ever freezes.	**A popsicle that only Danielle admires** ever freezes.
	sentential_negation_npi_licensor_present	64.1	90.7	77.0	90.8	86.3	87.6	85.9	82.0	90.5	**Teresa had not** ever sold a movie theater.	**Teresa had probably** ever sold a movie theater.
	sentential_negation_npi_scope	54.7	57.4	46.3	52.1	55.6	36.8	44.4	48.0	43.6	**Every son of Jerry who has insulted Jerry can not** ever die.	**Every son of Jerry who has not insulted Jerry can** ever die.
QUANTIFIERS	existential_there_quantifiers_1	72.3	85.1	73.3	82.7	79.0	79.5	85.5	83.9	77.2	There were no legislatures working hard.	There were most legislatures working hard.
	existential_there_quantifiers_2	40.9	42.4	56.2	63.2	42.9	41.6	47.5	29.6	36.2	All convertibles weren't there existing.	There weren't all convertibles existing.
	superlative_quantifiers_1	65.8	66.2	70.4	81.5	71.6	62.6	81.3	82.7	74.2	No girl attacked fewer than two waiters.	No girl attacked at most two waiters.
	superlative_quantifiers_2	57.6	55.8	54.0	59.3	58.3	53.6	58.9	60.4	55.7	The teenager does tour at most nine restaurants.	No teenager does tour at most nine restaurants.
SUBJECT VERB AGREEMENT	distractor_agreement_relational_noun	48.1	39.6	42.7	35.7	43.5	42.9	45.5	40.5	42.5	The sketch of those trucks **hasn't** hurt Alan.	The sketch of those trucks **haven't** hurt Alan.
	distractor_agreement_relative_clause	46.8	34.6	40.2	37.7	42.5	42.5	39.4	41.3	39.1	Boys that aren't disturbing Natalie **suffer**.	Boys that aren't disturbing Natalie **suffers**.
	irregular_plural_subject_verb_agreement_1	50.8	57.6	56.3	56.5	55.3	54.7	57.4	55.8	53.6	Those radii **have** scared that teenager.	Those radii **has** scared that teenager.
	irregular_plural_subject_verb_agreement_2	51.4	73.7	59.2	65.8	62.6	59.4	63.9	61.8	60.4	The **child** isn't attacking Becky.	The **children** isn't attacking Becky.
	regular_plural_subject_verb_agreement_1	53.9	71.6	71.4	67.5	67.7	62.0	67.3	60.2	63.0	Paula **references** Robert.	Paula **reference** Robert.
	regular_plural_subject_verb_agreement_2	53.6	69.2	59.4	59.7	55.0	56.2	58.5	60.7	60.2	The **associations** talk about Stacey.	The **association** talk about Stacey.

Figure 4: Performance of model on all 67 paradigms in BLiMP test suite along with examples of minimal pairs

of 1e-4, batch size of 16 and limit the maximum sequence length to 128. This model is trained for 10 epochs with max step size of 260. We train this on a single V100 GPU. To tokenize the words we use Byte Pair Encoder (BPE) (Gage, 1994) based tokenizer with vocabulary size set to 52,000 and minimum frequency set to 2. The rest of the hyperparameters are set to their default settings in the Transformers library (Wolf et al., 2019).

A.2 Comprehensive Results

Figure 4 illustrates the organization of the BLiMP test suite and the performance of all models along with examples of minimal pairs from each of the 67 paradigms.

A.2.1 Organization of BLiMP

BLiMP consists of 67 minimal pair paradigms grouped into 12 distinct linguistic phenomena: anaphor agreement, argument structure, binding, control/raising, determiner-noun agreement, ellipsis, filler gap, irregular forms, island effects, NPI licensing, quantifiers, and subject-verb agreement. Each paradigm comprises 1,000 sentence pairs in English and isolates specific phenomenon in syntax, morphology, or semantics. A complete description of each linguistic phenomenon and finer details of the test suite can be found in Warstadt et al. (2020).

A.2.2 Models

A total of 9 models are used in our study. (1) The Random Baseline model that is specifically trained such that it acquires no grammatical structure from the training data and only acquires English vocabulary (2) the Wikipedia-4 model that is trained on scripted ADS English data (3) the English ADS model that is trained on transcriptions of spontaneous, conversational speech in English (4) the English CDS model (5) the German CDS model (6) the French CDS model (7) the Polish CDS model (8) the Japanese CDS model (9) the Indonesian CDS model, where models 4 through 9 are trained on conversational CDS data from 6 different languages.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
8

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
8

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
3

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
A.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4, A.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5, 8

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
A.1

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.